# Modern Paradigms in Learning and Decision-Making
# Lecture #3: Omniprediction and Universal Adaptability*

Juan C. Perdomo†

Fall 2026

## 1 Introduction

In the last lecture, we introduced Outcome Indistinguishability and analyzed how to learn a predictor $\tilde{p}$ that is OI with respect to any set of tests $\mathcal{A}$ using a small amount of data. Later on in the course, we will introduce more sophisticated algorithms with faster runtimes and improved sample efficiency. However, in this class, we will take a detour and delve deeper into the connections between indistinguishability and notions of decision-making and loss minimization.

We start by recalling the OI implies loss minimization result from the last lecture:

**Proposition 1.1** ( [GHK+23]). *Fix any hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \to [0,1]\}$, error parameter $\varepsilon \geq 0$, and proper loss function $\ell$. Assume that $\tilde{p}$ is $\varepsilon$-OI with respect to the following class of functions*

$$A(x,p,y) = \ell(p,y) \text{ and } A_h(x,p,y) = \ell(h(x),y)$$

*for every $h \in \mathcal{H}$. Then,*

$$\mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\mathrm{Ber}(p^*(x))}} \ell(\tilde{p}(x), y^*) \leq \min_{h\in\mathcal{H}} \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\mathrm{Ber}(p^*(x))}} [\ell(h(x), y^*)] + 2\varepsilon.$$

In this lecture, we strengthen the connection between OI and loss minimization to find decision rules that are:

- Not only good for a fixed proper loss function $\ell$, but which can guarantee loss minimization with respect to many loss functions *simultaneously*.
- Not only good with respect to a fixed distribution $\mathcal{D}$, but also optimal with respect to many different distributions $\mathcal{D}'$.

We start by introducing the concept of a postprocessing function $\pi_\ell$.

*These notes are a work in progress and have not been subjected to the usual scrutiny reserved for formal publications.

**Definition 1.2.** *Fix $\mathcal{Y} = \{0, 1\}$. Let $\widehat{\mathcal{Y}}$ be the prediction or decision set and let $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$ be any loss function. We define the postprocessing function $\pi_\ell : [0, 1] \to \widehat{\mathcal{Y}}$ to be*

$$\pi_\ell(p) = \operatorname*{argmin}_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathop{\mathbf{E}}_{\tilde{y} \sim \mathrm{Ber}(p)} \ell(\widehat{y}, \tilde{y}).$$

The postprocessing function takes in a probability $p$ (specifying a conditional distribution over outcomes) and picks the best action for the loss function $\ell$ assuming that outcomes are sampled according to that conditional distribution.

Since the postprocessing $\pi_\ell$ chooses the best action for $\ell$ in this "simulated" world where outcomes are sampled according to $p$, the expected loss is better than that of any other decision rule or hypothesis $h$. For any $x \in \mathcal{X}$ and $h : \mathcal{X} \to \widehat{\mathcal{Y}}$:

$$\mathop{\mathbf{E}}_{\tilde{y} \sim \mathrm{Ber}(p)} \ell(\pi_\ell(p), \tilde{y}) \leq \mathop{\mathbf{E}}_{\tilde{y} \sim \mathrm{Ber}(p)} \ell(h(x), \tilde{y}).$$

Therefore, if we replace $p$ by a predictor $\tilde{p}(x)$ and take a distribution $\mathcal{D}_x$ over features $x \in \mathcal{X}$, we get that for any other decision rule $h$:

$$\mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} \ell(\pi_\ell(\tilde{p}(x)), \tilde{y}) \leq \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} \ell(h(x), \tilde{y}).$$

Using OI we will be able to translate this guarantee from the simulated world where outcomes are drawn from $\tilde{p}$ to the real world where they are drawn from $p^*$.

Before we do that, we remark that postprocessings are easy to compute. In particular, for Bernoulli outcomes $y$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\tilde{y} \sim \mathrm{Ber}(p)} \ell(\widehat{y}, \tilde{y}) &= \mathop{\mathbf{E}}_{\tilde{y} \sim \mathrm{Ber}(p)} [\ell(\widehat{y}, 1) \cdot \mathbf{1}\{\tilde{y} = 1\} + \ell(\widehat{y}, 0)(1 - \mathbf{1}\{\tilde{y} = 1\})] \\ &= \ell(\widehat{y}, 1) \cdot p + \ell(\widehat{y}, 0)(1 - p)\end{aligned}$$

Therefore, the postprocessing function can be computed by solving the following optimization problem that involves no randomness or expectations anywhere:

$$\operatorname*{argmin}_{\widehat{y} \in \widehat{\mathcal{Y}}} \ell(\widehat{y}, 1) \cdot p + \ell(\widehat{y}, 0)(1 - p).$$

It's helpful to see a few examples to convince ourselves of this point:

- *0-1 loss.* Here $\widehat{\mathcal{Y}} = \{0, 1\}$ and $\ell_{01}(\widehat{y}, y) = \mathbf{1}\{\widehat{y} \neq y\}$. Predicting $\widehat{y} = 1$ costs $1 - p$ and predicting $\widehat{y} = 0$ costs $p$, so the postprocessing is just a thresholding of $p$:

$$\pi_{\ell_{01}}(p) = \mathbf{1}\{p \geq 1/2\}.$$

- *Squared loss.* Here $\widehat{\mathcal{Y}} = [0, 1]$ and $\ell_2(\widehat{y}, y) = (\widehat{y} - y)^2$. The expected loss is $(\widehat{y} - 1)^2 \cdot p + \widehat{y}^2(1 - p)$, which by taking derivatives is minimized at

$$\pi_{\ell_2}(p) = p.$$

- *Cost-sensitive classification.* Here $\widehat{\mathcal{Y}} = \{0,1\}$ and, for a fixed $c \in (0,1)$, a false positive (where $\widehat{y} = 1$ and $y = 0$) costs $c$ while a false negative (where $\widehat{y} = 0$ and $y = 1$) costs $1 - c$:

$$\ell_c(\widehat{y}, y) = c \cdot \mathbf{1}\{\widehat{y} = 1, y = 0\} + (1-c) \cdot \mathbf{1}\{\widehat{y} = 0, y = 1\}.$$

  Predicting $\widehat{y} = 1$ costs $c(1-p)$ and predicting $\widehat{y} = 0$ costs $(1-c)p$, so

$$\pi_{\ell_c}(p) = \mathbf{1}\{p \geq c\}.$$

  The 0-1 loss is the special case where $c = 1/2$.

## 2 Omniprediction

Having introduced the idea of the postprocessing, we now define the concept of an omnipredictor.

**Definition 2.1** (Omnipredictor [GKR+22])**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0,1\}$, let $\mathcal{H}$ be a set of hypotheses or decision-rules, and let $\mathcal{L}$ be a set of loss functions $\ell$. A predictor $\tilde{p}$ is an $(\varepsilon, \mathcal{L}, \mathcal{H})$-omnipredictor if for every loss function $\ell \in \mathcal{L}$, the postprocessing function $\pi_\ell$ satisfies*

$$\mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}} \ell(\pi_\ell(\tilde{p}(x)), y^*) \leq \min_{h\in\mathcal{H}} \mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}} \ell(h(x), y^*) + \varepsilon.$$

In plain English, a predictor $\tilde{p}$ is an omnipredictor if we can use it to guarantee good decision-making with respect to many downstream loss functions. Given a loss function $\ell$, we can postprocess the outputs of the predictor to guarantee good decisions for that specific loss function $\ell$.

Omnipredictors enable *flexibility*. Normally in learning, we fix a loss function $\ell$ in advance, then we collect data from the distribution $\mathcal{D}$ and find a predictor or decision-rule that is optimal for that loss function. In omniprediction, we do not need to specify any loss function in advance. We summarize all of the necessary information from $\mathcal{D}$ into a simple predictor $\tilde{p}$ such that no matter the loss function we end up choosing, we can find an optimal decision rule for that loss without collecting more data or performing expensive computations.

Notice that to find the decision-rule, however, we allow ourselves to "tweak" $\tilde{p}$ via a cheap post-processing. This is one of the key methodological innovations of the omniprediction definition that allows us to move past prior "robust" notions of learning that aim to find predictors satisfying guarantees such as

$$\max_{\ell\in\mathcal{L}} \mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}} \ell(\tilde{p}(x), y^*) \leq \min_{h\in\mathcal{H}} \max_{\ell\in\mathcal{L}} \mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}} \ell(h(x), y^*) + \varepsilon.$$

In these versions, we are not allowed to postprocess $\tilde{p}$ but rather must have one model that minimizes the max loss in the set $\mathcal{L}$.

If $\mathcal{L}$ contains many different losses, then $\tilde{p}$ can only output trivial things since it needs to "split the difference" and balance between the competing objectives. By changing the order of quantifiers and allowing a postprocessing, we can find decisions rules which can do much better.

**Example 2.2.** *Let* $\widehat{\mathcal{Y}} = [0,1]$ *and* $\mathcal{L} = \{\ell_{\text{abs}}, \bar{\ell}_{\text{abs}}\}$*, where* $\ell_{\text{abs}}(\widehat{y}, y) = |\widehat{y} - y|$ *and* $\bar{\ell}_{\text{abs}} = 1 - \ell_{\text{abs}}$*. If we want to be good with respect to* $\max\{\ell_{\text{abs}}, \bar{\ell}_{\text{abs}}\}$*, we need to set* $\tilde{p}(x) = 1/2$*, regardless of what* $p^*(x)$ *is and*

$$\max_{\ell \in \mathcal{L}} \mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}} \ell(1/2, y^*) = 1/2.$$

*If we instead allow ourselves to postprocess,*

$$\pi_{\ell_{\text{abs}}}(p) = \mathbf{1}\{p \geq 1/2\} \text{ and } \pi_{\bar{\ell}_{\text{abs}}}(p) = \mathbf{1}\{p < 1/2\},$$

*we can do much better than a max loss of* $1/2$*. For* $\tilde{p} = p^*$*, both postprocessed rules have expected loss* $\min\{p^*(x), 1 - p^*(x)\}$ *at every* $x$*, so*

$$\max_{\ell \in \mathcal{L}} \mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}} \ell(\pi_\ell(p^*(x)), y^*) = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x} \min\{p^*(x), 1 - p^*(x)\} \leq 1/2,$$

*which is the smallest possible loss for each* $\ell \in \mathcal{L}$ *separately. The max loss is* $0$ *whenever outcomes are deterministic,* $p^*(x) \in \{0, 1\}$*.*

Notice the change in the quantifiers. In robust learning, we find a decision rule that is good for all losses. In omniprediction, we find one model such that for any loss, there exists a postprocessing of the predictor that achieves low loss.

## 2.1 Omniprediction via Outcome Indistinguishability

It is easy to show that if the conditional distribution over outcomes is equal to $p^*$, that is $y^* \sim \text{Ber}(p^*(x))$, then $p^*$ is itself an $(0, \mathcal{L}, \mathcal{H})$-omnipredictor with respect to any class of losses $\mathcal{L}$ and decision rules $\mathcal{H}$ (try it!).

The interesting fact is that we can find predictors $\tilde{p}$ that are very different from $p^*$ which still satisfy this property that we can treat their outputs $\tilde{p}(x)$ "as if" they were the true conditional distribution. In particular, we can extend the OI machinery we had developed in the previous lecture to derive indistinguishability conditions that imply omniprediction.

**Proposition 2.3.** *Fix any hypothesis class* $\mathcal{H}$*, error parameter* $\varepsilon \geq 0$*, and set of loss functions* $\mathcal{L}$*. Assume that* $\tilde{p}$ *is* $\varepsilon$*-OI with respect to the following tests*

$$A_\ell(x, p, y) = \ell(\pi_\ell(p), y) \text{ and } A_{\ell,h}(x, p, y) = \ell(h(x), y)$$

*for every* $h \in \mathcal{H}$ *and* $\ell \in \mathcal{L}$*. Then,* $\tilde{p}$ *is a* $(2\varepsilon, \mathcal{L}, \mathcal{H})$*-omnipredictor:*

$$\mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\text{Ber}(p^*(x))}} \ell(\pi_\ell(\tilde{p}(x)), y^*) \leq \min_{h\in\mathcal{H}} \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\text{Ber}(p^*(x))}} [\ell(h(x), y^*)] + 2\varepsilon.$$

*Proof.* From the first set of tests $A_\ell$ that depend only on $\ell$ we have that for any $\ell \in \mathcal{L}$:

$$\left| \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\text{Ber}(p^*(x))}} [\ell(\pi_\ell(\tilde{p}(x)), y^*)] - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\text{Ber}(\tilde{p}(x))}} [\ell(\pi_\ell(\tilde{p}(x)), \tilde{y})] \right| \leq \varepsilon. \tag{1}$$

From the second set of conditions that depend on $\ell$ and $h$ we have that for any pair $(\ell, h) \in (\mathcal{L}, \mathcal{H})$:

$$\left| \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\text{Ber}(p^*(x))}} [\ell(h(x), y^*)] - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\text{Ber}(\tilde{p}(x))}} [\ell(h(x), \tilde{y})] \right| \leq \varepsilon. \tag{2}$$

Now fix any loss function $\ell$ in the set $\mathcal{L}$. For any $h \in \mathcal{H}$,

$$\begin{aligned} \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\text{Ber}(p^*(x))}} \ell(\pi_\ell(\tilde{p}(x)), y^*) &\leq \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\text{Ber}(\tilde{p}(x))}} \ell(\pi_\ell(\tilde{p}(x)), \tilde{y}) + \varepsilon \\ &\leq \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\text{Ber}(\tilde{p}(x))}} \ell(h(x), \tilde{y}) + \varepsilon \\ &\leq \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\text{Ber}(p^*(x))}} \ell(h(x), y^*) + 2\varepsilon \end{aligned}$$

The first inequality follows from (1), the second inequality follows from the fact that $\pi_\ell$ is the postprocessing function and the last inequality follows from (2). Since these 3 inequalities are true for every $h \in \mathcal{H}$, they in particular hold for the best $h \in \mathcal{H}$. Since we picked $\ell \in \mathcal{L}$ to be arbitrary, it also holds uniformly over all $\ell \in \mathcal{L}$ hence $\tilde{p}$ must be an omnipredictor. □

Notice how general the proof is. We did not assume anything at all about the class of loss functions $\mathcal{L}$. They do not need to be convex, continuous, or satisfy any kind of regularity assumption. Furthermore, the class $\mathcal{H} \subseteq \{\mathcal{X} \to \widehat{\mathcal{Y}}\}$ and the set $\widehat{\mathcal{Y}}$ are again totally arbitrary.

The proof is also nearly identical to that of the previous proposition. The only thing we did was to *a*) expand the set of tests so that it works for many losses at the same time and *b*) slightly change the definition of the tests so that we no longer required the losses $\ell$ to be proper in the theorem statement. This, however, is only a cosmetic change. The underlying reason for why the argument works is still the same as per the following exercise:

**Exercise.** Let $\ell : \widehat{\mathcal{Y}} \times \{0, 1\} \to \mathbb{R}$ be any loss function. Prove that the loss function $\ell' : [0, 1] \times \{0, 1\} \to \mathbb{R}$ equal to $\ell'(p, y) = \ell(\pi_\ell(p), y)$ is a proper loss.

Another key advantage of the OI machinery is that, as mentioned previously, it enables us to treat the predictor $\tilde{p}$ "as if" it was the true conditional distribution $p^*$ even though $p^*$ and $\tilde{p}$ can be very different. As a proof technique, this allows us to learn predictors $\tilde{p}$ that are "as good" as $p^*$ for decision-making without depending on the complexity of $p^*$ for learning.

**Exercise.** Let $\mathcal{X} = \{-1, +1\}^2$ with $\mathcal{D}_x$ uniform over $\mathcal{X}$. Fix $\gamma, \beta > 0$ with $\gamma + \beta \leq 1/2$ and assume that

$$p^*(x) = \mathbf{Pr}[y = 1 \mid \mathcal{X} = x] = \frac{1}{2} + \gamma x_1 + \beta x_1 x_2.$$

Consider the predictor $\tilde{p}(x) = \frac{1}{2} + \gamma x_1$. Compute the total variation distance between $\tilde{p}$ and $p^*$:

$$\mathrm{TV}(\tilde{p}, p^*) = \mathop{\mathbf{E}}_{x \sim \mathcal{D}_x} |p^*(x) - \tilde{p}(x)|$$

Let $\mathcal{L}$ be the set of *all* loss functions $\ell : \widehat{\mathcal{Y}} \times \{0,1\} \to \mathbb{R}$, for any decision set $\widehat{\mathcal{Y}}$. Let $\mathcal{H}$ be the set of all decision rules that look at a single coordinate: $h(x) = g(x_1)$ or $h(x) = g(x_2)$ for some $g : \{-1,+1\} \to \widehat{\mathcal{Y}}$. Prove that $\tilde{p}$ is a $(0, \mathcal{L}, \mathcal{H})$-omnipredictor.

# 3 Universal Adaptability

So far, we have shown how to find decision rules that are good for many loss functions at the same time. In this section, we turn to the second goal from the introduction and show how to find decision rules that are good for many distributions at the same time.

As a motivating example, assume we train a risk predictor $\tilde{p}$ on data from patients from a specific hospital. We know that for the loss function $\ell$ we chose, that $\tilde{p}$ is approximately risk minimizing with respect to the distribution of patients we trained on. However, we would like to be able to use this predictor on different hospitals and guarantee that the predictor we trained is loss minimizing over *different* patient populations. We will illustrate a way to do just that.

Throughout, we assume that the conditional distribution of outcomes $p^*(x) = \mathbf{Pr}[y = 1 \mid \mathcal{X} = x]$ stays fixed, and that only the distribution over features $x$ changes. This is commonly referred to as *covariate shift*.

In the hospital example, this means that for any given feature $x$, the likelihood of the health outcome $y$ for individuals with features $x$, $\mathbf{Pr}_{\mathcal{D}}[y = 1 | \mathcal{X} = x]$, is the same in both populations, but the likelihood of seeing patients with a particular set of features is not the same in one hospital versus the other, $\mathbf{Pr}_{\mathcal{D}_x}[\mathcal{X} = x] \neq \mathbf{Pr}_{\mathcal{D}'_x}[\mathcal{X} = x]$.

The standard tool for dealing with covariate shift is importance weighting. If we know the target distribution $\mathcal{D}'_x$ that we want to do well on, then for any function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}'_x}[g(x)] = \mathop{\mathbf{E}}_{x \sim \mathcal{D}_x}[\omega(x) \cdot g(x)] \quad \text{where} \quad \omega(x) = \frac{\mathcal{D}'_x(x)}{\mathcal{D}_x(x)}.$$

Indeed,

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim \mathcal{D}'_x}[g(x)] &= \sum_{x \in \mathcal{X}} g(x) \mathop{\mathbf{Pr}}_{\mathcal{D}'_x}[\mathcal{X} = x] \\
&= \sum_{x \in \mathcal{X}} g(x) \mathop{\mathbf{Pr}}_{\mathcal{D}_x}[\mathcal{X} = x] \cdot \frac{\mathbf{Pr}_{\mathcal{D}'_x}[\mathcal{X} = x]}{\mathbf{Pr}_{\mathcal{D}_x}[\mathcal{X} = x]} \\
&= \sum_{x \in \mathcal{X}} g(x) \mathop{\mathbf{Pr}}_{\mathcal{D}_x}[\mathcal{X} = x] \cdot \omega(x) \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_x}[g(x)\omega(x)]
\end{aligned}$$

The function $\omega$ is often called the importance weight or the propensity score. Importance weighting allows us to take the expectation of a function $g$ over a distribution $\mathcal{D}$ and be able to evaluate it over a different distribution $\mathcal{D}'$. If we have a predictor $\tilde{p}$ and a loss $\ell$, we can evaluate how well it does on $\mathcal{D}'$ by computing

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_x, y^*\sim\mathrm{Ber}(p^*(x))} \ell(\tilde{p}(x), y^*)\omega(x) = \mathop{\mathbf{E}}_{x\sim\mathcal{D}'_x, y^*\sim\mathrm{Ber}(p^*(x))} \ell(\tilde{p}(x), y^*)$$

Importantly, we can only do this for distributions $\mathcal{D}'$ that lie in the support of $\mathcal{D}$. That is, if $\mathbf{Pr}_{\mathcal{D}'_x}[\mathcal{X} = x] > 0$ then it must be the case that $\mathbf{Pr}_{\mathcal{D}_x}[\mathcal{X} = x] > 0$. If $\mathcal{D}'$ places positive mass on data points we've never seen, then there is no way to reweight $\mathcal{D}$ into $\mathcal{D}'$. On a technical level, the derivation we did above breaks since

$$\frac{\mathbf{Pr}_{\mathcal{D}'_x}[\mathcal{X} = x]}{\mathbf{Pr}_{\mathcal{D}_x}[\mathcal{X} = x]}$$

is undefined since the denominator is 0.

A second requirement is that if we want to be good with respect to a distribution $\mathcal{D}'$ we need to know *in advance* what the right reweighting function $\omega$ is. Practically speaking, this is not always the case. Ideally, we could train a model once and deploy it everywhere without having to understand exactly what the reweighting could be. This is exactly the concept of *universal adaptability* which is yet another consequence of OI.

**Definition 3.1** (Universal Adaptability [KKG+22, KP23])**.** *Fix a distribution $\mathcal{D}$ and let $\mathcal{W}$ be a class of importance weight functions*

$$\mathcal{W} \subseteq \left\{\omega : \mathcal{X} \to \mathbb{R}_{\geq 0} \text{ where } \mathop{\mathbf{E}}_{\mathcal{D}_x}[\omega(x)] = 1\right\}.$$

*For each $\omega \in \mathcal{W}$, let $\mathcal{D}_\omega$ be the distribution over features obtained by reweighting $\mathcal{D}_x$ by $\omega$,*

$$\mathcal{D}_\omega(x) = \omega(x) \cdot \mathcal{D}_x(x).$$

*Let $\mathcal{H}$ be a set of decision rules, and $\ell$ be a proper loss function. A predictor $\tilde{p}$ is $(\varepsilon, \mathcal{W}, \mathcal{H})$-universally adaptable with respect to $\ell$ if for every $\omega \in \mathcal{W}$:*

$$\mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_\omega \\ y^*\sim\mathrm{Ber}(p^*(x))}} \ell(\tilde{p}(x), y^*) \leq \min_{h\in\mathcal{H}} \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_\omega \\ y^*\sim\mathrm{Ber}(p^*(x))}} \ell(h(x), y^*) + \varepsilon.$$

Notice the requirement that the importance weight function averages to 1 over the distribution $\mathcal{D}_x$. This is exactly what is needed for $\mathcal{D}_\omega$ to be a probability distribution. Since $\mathbf{Pr}_{\mathcal{D}_\omega}[\mathcal{X} = x] = \omega(x) \cdot \mathbf{Pr}_{\mathcal{D}_x}[\mathcal{X} = x]$,

$$\sum_{x\in\mathcal{X}} \mathop{\mathbf{Pr}}_{\mathcal{D}_\omega}[\mathcal{X} = x] = \sum_{x\in\mathcal{X}} \omega(x) \cdot \mathop{\mathbf{Pr}}_{\mathcal{D}_x}[\mathcal{X} = x] = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x}[\omega(x)] = 1.$$

We have stated the definition for a single proper loss $\ell$ only to keep things simple. One could just as well define an omniprediction version of universal adaptability, where $\tilde{p}$ is required to be an $(\varepsilon, \mathcal{L}, \mathcal{H})$-omnipredictor over every $\mathcal{D}_\omega$. Everything we say below carries over to that setting by using the tests from Proposition 2.3 in place of those from Proposition 1.1.

Universal adaptability is again a consequence of OI. The tests are the same as in Proposition 1.1, except that we multiply each one by an importance weight.

**Proposition 3.2** ( [KKG+22,KP23]). *Fix any hypothesis class $\mathcal{H}$, proper loss $\ell$, class of importance weights $\mathcal{W}$, and error parameter $\varepsilon \geq 0$. Assume that $\tilde{p}$ is $\varepsilon$-OI over $\mathcal{D}$ with respect to the tests*

$$A_\omega(x, p, y) = \omega(x) \cdot \ell(p, y) \text{ and } A_{h,\omega}(x, p, y) = \omega(x) \cdot \ell(h(x), y)$$

*for every $h \in \mathcal{H}$ and $\omega \in \mathcal{W}$. Then, $\tilde{p}$ is $(2\varepsilon, \mathcal{W}, \mathcal{H})$-universally adaptable with respect to $\ell$.*

To see why this is true, fix any $\omega \in \mathcal{W}$ and $h \in \mathcal{H}$. By the importance weighting identity applied once to the function $g(x) = \mathbf{E}_{y^* \sim \mathrm{Ber}(p^*(x))} \ell(h(x), y^*)$ and once to $g(x) = \mathbf{E}_{\tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))} \ell(h(x), \tilde{y})$, we have that

$$\begin{aligned} &\mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_\omega \\ y^* \sim \mathrm{Ber}(p^*(x))}} \ell(h(x), y^*) - \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_\omega \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} \ell(h(x), \tilde{y}) \\ &\quad = \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} \omega(x) \cdot \ell(h(x), y^*) - \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} \omega(x) \cdot \ell(h(x), \tilde{y}). \end{aligned}$$

At this point, we've seen this proof template enough times that you can work out the details yourself as an exercise. Notice that the learner needs to know the class $\mathcal{W}$ at training time, in order to write down the tests. However, they do not need to know *which* $\omega \in \mathcal{W}$ describes the population that $\tilde{p}$ will eventually be deployed on. This is exactly the "train once, deploy everywhere" guarantee we wanted to achieve.

**The price of universal adaptability.** Recall that the BOOST procedure and the sample complexity bound from the last lecture required the discrete derivatives of the tests to be bounded,

$$\sup_{A \in \mathcal{A}} |\Delta A(x, p)| \leq 1.$$

For the tests in Proposition 1.1, this condition holds as long as the loss $\ell$ takes values in $[0, 1]$, since the discrete derivative of $A_h$ is

$$\Delta A_h(x, p) = \ell(h(x), 1) - \ell(h(x), 0) \in [-1, 1].$$

For the weighted tests, the discrete derivative picks up a factor of $\omega(x)$,

$$\Delta A_{h,\omega}(x, p) = \omega(x) \cdot \big(\ell(h(x), 1) - \ell(h(x), 0)\big),$$

which is only bounded by the largest importance weight in the class,

$$\omega_{\max} = \max_{\omega \in \mathcal{W}} \max_{x \in \mathcal{X}} \frac{\mathbf{Pr}_{\mathcal{D}_\omega}[\mathcal{X} = x]}{\mathbf{Pr}_{\mathcal{D}_x}[\mathcal{X} = x]}.$$

To apply the results from the last lecture, we need to rescale things. Define $A'_{h,\omega} = A_{h,\omega}/\omega_{\max}$ and $A'_\omega = A_\omega/\omega_{\max}$, so that the discrete derivatives of the rescaled tests are again bounded by 1. Since the OI error is linear in the test, being $\varepsilon'$-OI with respect to $A'_{h,\omega}$ is the same as being $\varepsilon' \cdot \omega_{\max}$-OI

with respect to $A_{h,\omega}$:

$$\left| \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\mathrm{Ber}(p^*(x))}} A_{h,\omega}(x,\tilde{p}(x),y^*) - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\mathrm{Ber}(\tilde{p}(x))}} A_{h,\omega}(x,\tilde{p}(x),\tilde{y}) \right|$$

$$= \omega_{\max} \cdot \left| \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\mathrm{Ber}(p^*(x))}} A'_{h,\omega}(x,\tilde{p}(x),y^*) - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\mathrm{Ber}(\tilde{p}(x))}} A'_{h,\omega}(x,\tilde{p}(x),\tilde{y}) \right|.$$

Consequently, to guarantee $\varepsilon$-OI with respect to the original tests $A_{h,\omega}$, we run BOOST on the rescaled tests with error parameter $\varepsilon' = \varepsilon/\omega_{\max}$. Plugging $\varepsilon'$ into the guarantees from the last lecture, the number of rounds of boosting goes from $1/\varepsilon^2$ to $\omega_{\max}^2/\varepsilon^2$, and the number of samples needed per round goes from $\log(|\mathcal{A}|/\delta)/\varepsilon^2$ to $\omega_{\max}^2 \log(|\mathcal{A}|/\delta)/\varepsilon^2$.

The dependence on $\omega_{\max}$ is natural. If a target population $\mathcal{D}_\omega$ places most of its mass on individuals that are rare under $\mathcal{D}_x$, then a sample from $\mathcal{D}_x$ contains very few of the individuals we care about, and we need a larger sample to say anything about them. For instance, if a particular data point is 10 times more likely to appear in $\mathcal{D}_\omega$ than in $\mathcal{D}_x$ then the max density ratio $\omega_{\max} \geq 10$ and we need 100 times more data in order to make sure that we see those rare members of the population.

**Exercise (Learning from $k$ hospitals).** Let's return to the hospital example from the beginning of the section. Suppose that we pool data from $k$ hospitals, where a fraction $\lambda_i > 0$ of the data comes from hospital $i$. Let $\mathcal{D}_i$ be the distribution over patients at hospital $i$, and assume that the features $x$ record which hospital a patient comes from, so that $c_i(x) = \mathbf{1}\{x \text{ comes from hospital } i\}$ is a function of $x$. The pooled distribution over features is the mixture

$$\mathcal{D}_x = \sum_{i=1}^{k} \lambda_i \mathcal{D}_i.$$

1. Find a class of $k$ importance weights $\mathcal{W} = \{\omega_1, \ldots, \omega_k\}$ such that $\mathcal{D}_{\omega_i} = \mathcal{D}_i$ for every $i$. Conclude that if $\tilde{p}$ is $(\varepsilon, \mathcal{W}, \mathcal{H})$-universally adaptable with respect to $\ell$, then $\tilde{p}$ is $\varepsilon$ loss minimizing with respect to $\mathcal{H}$ at *every* hospital.

2. Write down the tests from Proposition 3.2 for this class $\mathcal{W}$. How do they compare to the tests in Proposition 1.1?

3. Compute $\omega_{\max}$ for this class. Compared to learning a predictor that is only loss minimizing over the pooled distribution $\mathcal{D}_x$, how many more samples do we need to learn a universally adaptable one? Which hospital determines the blowup?

4. Let $\mathcal{D}_m = \sum_{i=1}^{k} \mu_i \mathcal{D}_i$ be any other mixture of the $k$ hospitals, where $\mu_i \geq 0$ and $\sum_i \mu_i = 1$. Show that $\tilde{p}$ is also $\varepsilon$ loss minimizing with respect to $\mathcal{H}$ over $\mathcal{D}_m$. That is, $\tilde{p}$ comes with guarantees for any new hospital network formed out of the original $k$ hospitals, regardless of how the patients are split between them.

## Bibliographic Notes

Omnipredictors were introduced by Gopalan, Kalai, Reingold, Sharan, and Wieder [GKR+22], who showed that a predictor that is multicalibrated with respect to $\mathcal{H}$ is an omnipredictor for the class of all convex, Lipschitz losses. Their proof goes through multicalibration [HKRR18] directly. The presentation in this lecture, where omniprediction follows from OI with respect to a set of tests built from the losses and the hypotheses, is due to Gopalan, Hu, Kim, Reingold, and Wieder [GHK+23]. That paper also shows how to instantiate the tests for specific classes of losses, and how the resulting conditions relate to calibration and multiaccuracy.

Universal adaptability was introduced by Kim, Kern, Goldwasser, Kreuter, and Reingold [KKG+22] in the context of statistical estimation. There, the goal is to estimate the mean of an outcome over a target population using data from a different source population, and the main result is that a multiaccurate predictor competes with propensity scoring without knowing the target population in advance. The formulation in terms of loss minimization over a class of reweighted distributions, and its combination with omniprediction follows Kim and Perdomo [KP23] who study these questions in the setting where predictions influence outcomes. Both works build on the OI framework of Dwork, Kim, Reingold, Rothblum, and Yona [DKR+21].

## References


[DKR+21] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1095–1108, 2021.

[GHK+23] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS)*, 2023.

[GKR+22] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS)*, 2022.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1939–1948, 2018.

[KKG+22] Michael P. Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

[KP23] Michael P. Kim and Juan C. Perdomo. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference (ITCS)*, 2023.